DEREVERBRATION

The *SGMSE* (Score-based Generative Model for Speech Enhancement) was used to recover clean speech from reverberant recordings using a diffusion-based approach. The *NCSN++* architecture was employed as the backbone, and the *OU-VE* (Ornstein–Uhlenbeck Variance Exploding) stochastic differential equation was used to model the forward noise injection and reverse denoising processes. By learning these dynamics over time, the model gradually removes noise from the input and produces high-quality clean speech.

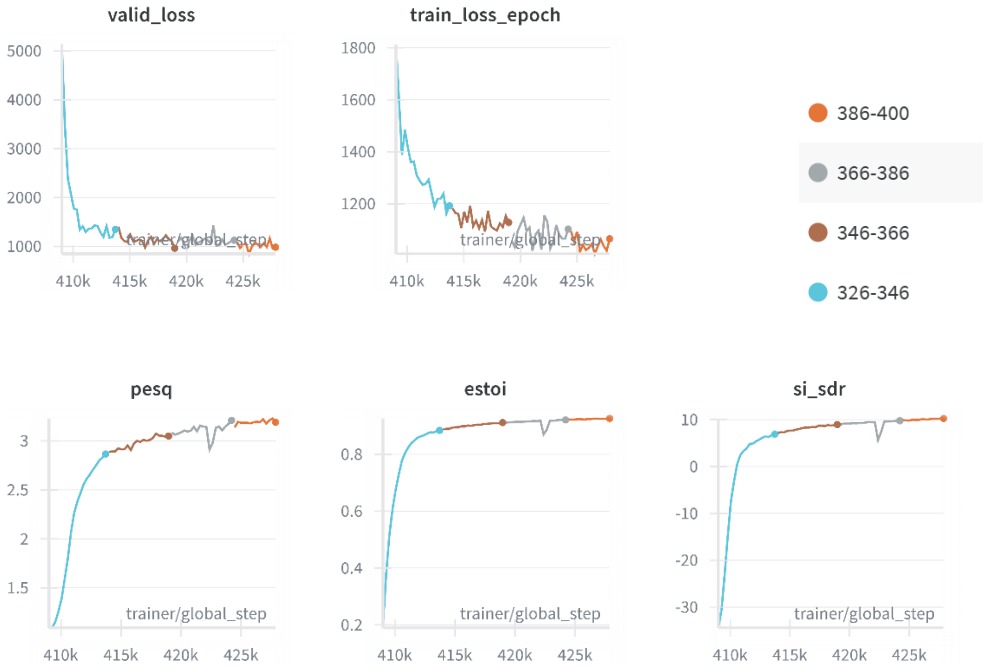
Datasets Used - ARNI RIRs, LibriSpeech Train-clean-360 (training) and Test-clean (evaluation).

METHODOLOGY

Note: Due to hardware constraints (limited GPU vRAM on Kaggle and Insufficient GPU resources in local hardware), processing the full dataset was not feasible. Therefore, a subset of the dataset and the IR Arni subset (0–25), was utilized for generating clean and noisy files for training.

Dataset split: train(2096) / validation(262) / test (262)

The model was fine-tuned in three sequential phases of 20 epochs each, and a final run with 14 epochs. The checkpoint from the last epoch of the final phase was used for generating enhanced speech and calculating DNSMOS scores.



Performance Metrics

Epoch Range	Best SI-SDR (dB)	Best eSTOI	Best PESQ	Validation Loss (Lowest)	Training Loss (Lowest)
326–346	6.87	0.884	2.87	1179.44	1162.32
346–366	8.95	0.912	3.08	962.95	1092.32
366–386	9.77	0.922	3.21	955.27	1018.83
386–400	10.26	0.926	3.23	849.43	1007.66

EVALUATION METRIC : DNSMOS

The checkpoint from the last epoch of the final fine-tuning phase was used to generate enhanced speech and calculate DNSMOS scores on a subset of the test dataset.

Model	Average SIG	Average BAK	Average OVRL
Sgmse_base	3.524	2.698	2.253
Improved (epoch- 365 cpkt)	**3.633**	**3.293**	**3.532**
Higher epoch run (epoch – 399 cpkt)	3.596	3.204	3.394

INFERENCE

Using SGMSE with NCSN++ (OU-VE), the higher-epoch run scores lower on DNSMOS because the model overfits the training noise and becomes overconfident. This leads to over-smoothed or unnatural backgrounds reducing BAK and OVRL. Thus, perceptual quality peaks at an intermediate epoch rather than full convergence.

RESULTS

DNSMOS SCORE

Sgmse_base - **2.25**

Improved (epoch- 365 cpkt) - **3.53**

The improved model has better DNSMOS overall (OVRL) scores than the base model, showing a significant increase from **2.25 to 3.53,** indicating improvement in perceived speech quality and listening experience.

The model can be improved by using early stopping based on perceptual metrics and training on a more diverse, larger dataset (which could not be done due to hardware constraints).